Datasheet:

Data Sources & Reasoning:

Population Data (CSV - Census)

For the population data set, we downloaded a CSV file from the United States Census. The Census dataset was created for the purpose of measuring the population size of particular cities, states, and the United States, over time. It is funded by the US federal government and helps keep tabs on the growing populations. It is also necessary to determine the number of seats each state has in the House of Representatives. The US census data was collected by sending letters out to all residents, who fill out their personal data. The accuracy of this data can be slightly off because of residents who have recently moved or people who have multiple residencies.

The biggest problem we came across was with this population dataset. We needed to download it and convert it to a CSV to put it into a readable format, but for some reason this document would not work in Google Colab, so we needed to use one computer at a time and e-mail a zip file back and forth to each other. The file can now be accessed as an excel sheet where each row is a separate city and each column contains one day each year from 2010-2017 that gives the population size. The dataset only provides a sample because it gives the population of each city for a single day each year. The data was not necessarily raw, but we had to search through it to get the specific cities and dates we wanted to get the correct population sizes over time.

The only people directly involved in this data collection were those who filled in their own information. These US citizens which were having their data collected were aware of it, as they voluntarily filled out the information which was sent to them.

Zillow Rental Index (Zillow - Quandl):

For our city rental price indices, we used Quandl to access their Zillow Rental Index dataset. Zillow is a popular online real estate discovery site which tracks home rental prices, sale prices, and property values. This dataset was created from the data that Zillow already tracks and collects. Zillow typically uses regulatory filings like building permits and home deeds to track changes in homes and their values. Thus, this set likely does not take into account any “special” instance housing, like low-income housing developments or transactions done privately (i.e. renting to family “under the table”). Very little preprocessing was done other than extracting the relevant years from the Quandl API.

The Rental Index includes a month-by-month breakdown of the average rental price (in US dollars) for homes in each city. We ultimately decided upon rental prices since this is a more instantaneous measure of property supply and demand. We reason that home values (the price that homes are typically selling for) are more inelastic than home rental prices. In other words, the factors that we are exploring will have a more immediate impact on rental prices than on home sale prices.

Unemployment Rate & Median Household Income (FRED - Quandl)

The FRED unemployment rate dataset was created for the advisement of the Federal Reserve Bank president on matters of economic policy. It was also created to promote economic education and enhance economic research as this data is accessible to the public. The Research Division of the Federal Reserve Bank of St. Louis is responsible for collecting the data. FRASER (Federal Reserve Archival System for Economic Research) is the digital database where all FRED economic data is stored. The St. Louis Fed worked with the US Government Printing Office and the Federal Depository Library Program to make this data accessible.

The Final Write Up document has a detailed explanation of how the Quandl data was accessed and parsed.

Crime Rate (FBI):

Our Crime Data comes from the Federal Bureau of Investigation's Uniform Crime Reporting Statistics Tool (ucrdatatool.gov). The Uniform Crime Reporting Program gathers crime statistics from law enforcement agencies throughout the United States, compiling crime reports into two overall crime rates: a violent crime rate and property crime rate. Violent crimes include incidents like murders, assaults, and rapes. Property crimes mostly include instances of theft or burglary.

There is likely little bias in this dataset since it comes from a government agency, but it certainly does not account for unreported crimes. This has a heavier impact on the "violent crimes" category since many incidents in this category, like sexual assault and rape, are notoriously underreported.

For preprocessing, we downloaded the web page for each city and used a web scraper to extract all of the yearly "crime rate" data from the web page. This data spans the years 1985 through 2014.